View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by PUblication MAnagement

# Evaluation of Natural Language Tools for Italian: EVALITA 2007

**B. Magnini[1], A. Cappelli[2], F. Tamburini[3], C. Bosco[4], A. Mazzei[4], V. Lombardo[4], F. Bertagna[5], N. Calzolari[5], A. Toral[5], V. Bartalesi Lenzi[2], R. Sprugnoli[2], and M. Speranza[1]**

[1] FBK-ricerca scientifica e tecnologica, Povo (Trento), Italy
[2] CELCT, Povo (Trento), Italy
[3] Dipartimento di Studi Linguistici e Orientali, Università di Bologna, Italy
[4] Dipartimento di Informatica, Università di Torino, Italy
[5] Istituto di Linguistica Computazionale, Consiglio Nazionale delle Ricerche, Pisa, Italy

E-mail: {magnini,speranza}@fbk.eu, {cappelli,bartalesi,sprugnoli}@celct.it, fabio.tamburini@unibo.it, {bosco,mazzei,lombardo}@di.unito.it, {bertagna,calzolari,toral}@ilc.cnr.it

**Abstract**

EVALITA 2007, the first edition of the initiative devoted to the evaluation of Natural Language Processing tools for Italian, provided a shared framework where participants' systems had the possibility to be evaluated on five different tasks, namely Part of Speech Tagging (organised by the University of Bologna), Parsing (organised by the University of Torino), Word Sense Disambiguation (organised by CNR-ILC, Pisa), Temporal Expression Recognition and Normalization (organised by CELCT, Trento), and Named Entity Recognition (organised by FBK, Trento).
We believe that the diffusion of shared tasks and shared evaluation practices is a crucial step towards the development of resources and tools for Natural Language Processing. Experiences of this kind, in fact, are a valuable contribution to the validation of existing models and data, allowing for consistent comparisons among approaches and among representation schemes. The good response obtained by EVALITA, both in the number of participants and in the quality of results, showed that pursuing such goals is feasible not only for English, but also for other languages.

## 1. Introduction

In the last decade, increasing emphasis has been given to the evaluation of newly developed techniques in Natural Language Processing. Evaluation per se, however, is not as useful for enhancing progress in the field as is the possibility of comparing results of different systems. In this perspective, the aim of the EVALITA initiative is to promote the development of language technologies for the Italian language, by providing a shared framework to evaluate different systems and approaches in a consistent manner.
A series of international evaluation campaigns have been organised recently, which propose tasks both for English and for other languages, sometimes including Italian. Among them are CoNLL[1] as far as Parsing and Named Entity Recognition are concerned, Senseval/Semeval (including Italian lexical sample)[2] for Word Sense Disambiguation, ACE program[3] (in particular Entity Detection and Recognition and Temporal Expression Recognition and Normalization), and finally CLEF[4] WiQA[5] and GeoCLEF[6] for Information Retrieval and Question Answering. Similarly to what had already been done for French with EASY[7] and for Portuguese with HAREM[8], EVALITA concentrates specifically on one single language, i.e. Italian.
Organized on a fully voluntary basis, EVALITA 2007 aimed at systematically proposing standards for Italian in some specific tasks where it was possible to exploit annotated material already available. These tasks were: Part of Speech Tagging (POS), Parsing (PAR), Word Sense Disambiguation (WSD), Temporal Expression Recognition and Normalization (TERN), and Named Entity Recognition (NER). As with the evaluation campaigns mentioned above, participants were provided with training data and had the chance to test their systems with the evaluation metrics and procedures to be used in the formal evaluation (Magnini & Cappelli, 2007).
For EVALITA 2007, we received a total number of 55 expressions of interest for the five tasks. In the end, 30 participants actually submitted their results, with the following distribution: 11 for POS, 8 for PAR, 1 for WSD, 4 for TERN, and 6 for NER. As shown in Table 1, four participants took part in more than one task. Overall, we had 21 different organizations; among them, eight were not Italian (i.e. Indian Institute of Information Technology, Linguistic Data Consortium, University of Alicante, University of Dortmund, University of Duisburg-Essen, University of Stuttgart-IMS, University of Pennsylvania and Yahoo! Research) and two were not academic (i.e. Yahoo! Research and Synthema). These more than satisfactory results make us think that it will be worth to work towards making EVALITA become a regular event

[1] http://ifarm.nl/signll/conll/
[2] http://www.senseval.org/
[3] http://www.nist.gov/speech/tests/ace/
[4] http://www.clef-campaign.org/
[5] http://ilps.science.uva.nl/WiQA/
[6] http://ir.shef.ac.uk/geoclef/
[7] http://www.limsi.fr/RS2005/chm/lir/lir11/
[8] http://www.linguateca.pt/HAREM/

(i.e. trying to organise an evaluation campaign for Italian every two years).

| Participant | Task | Institution(s), Country |
|---|---|---|
| FBKirst_Negri | TERN | FBK, Trento, IT |
| FBKirst_Pianta | PAR | |
| FBKirst_Zanoli | POS | |
| | NER | |
| IIIT_Mannem | PAR | IIIT, Hyderabad, IN |
| ILCcnrUniPi_Lenci | POS | ILC-CNR & Univ. Pisa, IT |
| LDC_Walker | NER | LDC, Philadelphia, USA |
| UniAli_Kozareva | NER | Univ. Alicante, ES |
| UniAli_Puchol | TERN | |
| UniAli_Saquete | TERN | |
| UniBa_Basile | WSD | Univ. Bari, IT |
| UniBoCilta_Romagnoli | POS | Univ. Bologna, IT |
| UniBoDslo_Tamburini | POS | |
| UniDort_Jungermann | NER | Univ. Dortmund, DE |
| UniDuE_Roessler | NER | Univ. Duisburg-Essen, DE |
| UniNa_Corazza | PAR | Univ. Napoli, IT |
| UniPg_Faina | TERN | Univ. Perugia, IT |
| UniPi_Attardi | PAR | Univ. Pisa, IT |
| UniPiSynthema_Deha | POS | Univ. Pisa & Synthema, IT |
| UniRoma1_Bos | POS | Univ. Roma La Sapienza, IT |
| UniRoma2_Zanzotto | PAR | Univ. Roma Tor Vergata, IT |
| UniStuttIMS_Schiehlen | POS | IMS – Univ. Stuttgart, DE |
| | PAR | |
| UniTn_Baroni | POS | Univ. Trento, IT |
| UniTo_Lesmo | POS | Univ. Torino, IT |
| | PAR | |
| UniVe_Delmonte | POS | Univ. Venezia, IT |
| UPenn_Champollion | PAR | Univ. Pennsylvania, USA |
| Yahoo_Ciaramita | POS | Yahoo!, Barcelona, ES |
| | NER | |

Table 1: List of participants to EVALITA 2007.

## 2. The Part of Speech Tagging Task

One of the tasks inside EVALITA 2007 was devoted to the evaluation of Part-of-Speech (PoS) taggers. As in other evaluation campaigns, the organisation provided a common framework for the evaluation of tagging systems in a consistent way, supplying the participants with manually annotated data as well as a scoring program for developing and evaluating their systems.
Eleven systems completed all the steps in the evaluation procedure and their outputs were officially submitted for this task by their developers.

### 2.1. Data description

The data sets were composed of various documents belonging mainly to journalistic and narrative genres, with small sections containing academic and legal/administrative prose. Two separate data sets were provided: the Development Set (DS), composed of 133,756 tokens, was used for system development and for the training phase, while a Test Set (TS), composed of 17,313 tokens, was used as a gold standard for systems evaluation. The ratio between DS and TS is 8/1.
These data have been manually annotated assigning to each token its lexical category (PoS-tag) with respect to two different tagsets producing two different subtasks.
The task organisation did not distribute any lexicon resource with EVALITA data. Each participant was allowed to use any available resource or could freely induce it from the training data.

### 2.2 Tagsets

The PoS-Tagging Task involved two different tagsets, used to classify the DS data and to be used to annotate TS data.
The structure and the principles underlying the tagset design are crucial, both for a coherent approach to lexical classification and to obtain better performance results with automatic techniques, thus they deserve a further discussion. Italian is one of the languages for which a set of annotation guidelines has been developed in the context of the EAGLES project (Monachini, 1995). Several research groups have been working on PoS annotation to develop Italian treebanks, such as VIT (Venice Italian Treebank – Delmonte, 2004) and TUT (Turin University Treebank – Bosco et al., 2000) and morphological analysers such as the one by XEROX. A comparison of the tagsets used by these groups with EAGLES guidelines reveals that, although there is general agreement on the main parts of speech to be used, considerable divergence exists as regards the actual classification of Italian words with respect to them. This is the main problematic issue, reflected also in the considerable classification differences operated by the Italian dictionaries.
For the reasons briefly outlined above, we decided to propose two different subtasks for the PoS-tagging evaluation campaign, the first using a traditional tagset (EAGLES-like), the second using a structurally different tagset (DISTRIB). We refer to the task guidelines (Tamburini & Seidenari, 2007) for an in-depth discussion of the two proposed tagsets.

### 2.3 Tokenisation issues

The problem of text segmentation (tokenisation) is a central issue in PoS-taggers comparison and evaluation. In principle every system could apply different tokenization rules leading to different outputs. In this first evaluation campaign we did not have the possibility of handling different tokenisation schemas and following the complex realignment work proposed, for example, inside the GRACE evaluation project (Adda et al., 1998). All the development and test data were provided in tokenised format. Participants were required to return the test set using the same tokenisation format, containing exactly the same number of tokens.

### 2.4 Evaluation Metrics

The evaluation was performed evaluating only the systems' outputs. The evaluation metrics were based on a token-by-token comparison and only one tag was allowed for each token. The considered metrics were:
a) *Tagging Accuracy*, defined as the number of correct

PoS-tag assignments divided by the total number of tokens in TS.

b) *Unknown Words Tagging Accuracy*, defined as the Tagging Accuracy restricting the computation to unknown words. In this context, for "unknown word" we meant a token present in TS but not in the DS. This metric allowed a finer evaluation on the most fruitful morphological techniques or heuristics used to manage unknown words for Italian, a typical challenging problem for automatic taggers.

### 2.5 Results and Discussion

Table 2 shows the global results of the EVALITA 2007 PoS Tagging Task for both tagsets, displaying systems' performances with respect to the proposed metrics.

A baseline algorithm, that assigns the most frequent tag for each known word and the absolute most frequent tag for unknown words, and some well known freely-available PoS-taggers (Brants TnT, 2000; Brill TBL tagger, 1994; Ratnaparkhi Maximum Entropy tagger, 1996; Daelemans et al. Memory Based tagger, 1996) have been inserted into the evaluation campaign as references for comparison purposes. All these taggers were tested by the organisers using the standard configurations described in the respective documentations.

| SYSTEM | EAGLES-like | | DISTRIB | |
|---|---|---|---|---|
| | TA | UWTA | TA | UWTA |
| Baseline | 90.43 | 32.96 | 89.48 | 43.06 |
| MXPOST | 96.14 | 86.50 | 95.15 | 86.65 |
| TnT | 96.82 | 86.73 | 95.96 | 86.80 |
| Brill | 94.39 | 58.90 | 94.13 | 60.71 |
| MBT | 95.48 | 77.53 | 95.02 | 78.13 |
| FBKirst_Zanoli | 98.04 | 95.02 | 97.68 | 94.65 |
| ILCcnrUniPi_Lenci | 97.65 | 94.12 | 96.70 | 93.14 |
| UniBoCILTA_Romagnoli | 96.79 | 91.48 | 94.80 | 90.72 |
| UniBoDSLO_Tamburini | 97.59 | 92.16 | 97.31 | 92.99 |
| UniRoma1_Bos | 96.76 | 87.41 | 96.21 | 88.69 |
| UniStuttIMS_Schielen | 97.15 | 89.29 | 97.07 | 92.23 |
| UniTn_Baroni | 97.89 | 94.34 | 97.37 | 94.12 |
| UniVe_Delmonte | 91.85 | 84.46 | 91.42 | 86.80 |
| Yahoo_Ciaramita_s1 | 96.78 | 87.78 | 96.61 | 88.24 |
| Yahoo_Ciaramita_s2 | 95.27 | 81.83 | 95.11 | 84.16 |
| UniPiSynthema_Deha | 88.71 | 79.49 | – | – |
| UniTo_Lesmo | 94.69 | 87.33 | – | – |

Table 2: Reference systems and participants' results with respect to Tagging Accuracy (TA) and UnknownWords Tagging Accuracy (UWTA).

Examining the systems' performances with respect to their structural features depicted in Table 2, we can make some tentative observations:

- there is a group of five systems that performs slightly better that the others exhibiting very high scores (97–98% of Tagging Accuracy), near to the state-of-the art performances obtained for English, a language on which there is a long tradition of studies for PoS automatic labelling;
- regarding the core methods implemented by the participants, Support Vector Machines seems to perform quite well: both systems using them are in the top five; the same observation holds for the systems obtained combining or stacking different taggers;
- additional lexical resources seems to play a major role in improving the performances: the systems employing morphological analyzers based on big lexica and special techniques for unknown word handling reached the top rankings. These results were clear when analyzing the scores considering the UnknownWords Tagging Accuracy metric;
- TnT obtains the best results among the considered reference systems: it embodies a standard, though well optimised, second-order HMM method and employs a sophisticated suffix analysis system that, even in absence of a lexical resource, produces good results;
- the performances obtained by the participating systems remained quite stable when changing the tagset: the best systems tend to exhibit a lowering in performances less than 0.5% when applied to the DISTRIB tagset.

## 3. The Parsing Task

The Penn Treebank has played an invaluable role in enabling the development of state-of-the-art parsing systems, but the strong focalization on it has left open several questions on parsers' portability. While strong empirical evidences demonstrate that results obtained on a particular treebank are unportable on other corpora (Gildea, 2001; Collins et al., 1999; Corazza et al., 2004), the validation of existing parsing models depends on the possibility of generalizing their results on corpora other than those on which they have been trained, tuned and tested.

The aim of the EVALITA 2007 Parsing Task, is to assess the current state-of-the-art in parsing Italian by encouraging the application of existing systems to this language, and to contribute to the investigation on the causes of this irreproducibility with reference to parsing models and treebank annotation schemes. It allowed to focus on Italian by exploring both different paradigms, i.e. constituency and dependency, and different approaches, i.e. rule-based and statistical.

The task consists in the activity of assigning a syntactic structure to a given Italian PoS tagged sentence, using a fully automatic parser and according to the annotation scheme of the development set, which can be selected between a dependency-based and a constituency-based one. It includes in fact two subtasks (dependency parsing and constituency parsing) with separate development datasets and evaluations.

### 3.1. Data description and evaluation metrics

The development data consisted of 2,000 sentences (i.e. about 58,000 annotated tokens) from the Turin University

Treebank (TUT[9]).
The corpus annotated in this treebank is organized in two subcorpora of one thousand sentences each, i.e. the Italian newspaper and the Italian legal Code.
The sentences are annotated respectively in TUT and in TUT-Penn format for the dependency and constituency parsing subtasks. For dependency, TUT implements a pure dependency annotation schema based on a rich set of grammatical relations, that also includes null elements in order to represent discontinuous and elliptical structures. For constituency, TUT-Penn adopts a Penn-like annotation, which has been produced by automatic conversion of TUT data, and that differentiates from Penn mainly because of the PoS tagset.
The evaluation of results is performed separately for dependency and constituency. For dependency results it is based on the three CoNLL standard metrics (Nivre et al., 2007):

- Labeled Attachment Score (LAS), the percentage of tokens with correct head and relation label;
- Unlabeled Attachment Score (UAS), the percentage of tokens with correct head;
- Label Accuracy (LAS2), the percentage of tokens with correct relation label.

For constituency, the evaluation is instead based on standard PARSEVAL measures:

- Brackets Precision (Br-P), the percentage of found brackets which are correct;
- Brackets Recall (Br-R), the percentage of brackets correct which are found;
- Brackets F (Br-F), the composition of the previous two measures that can be calculated by the following formula: 2 * (P * R) / (P + R).

### 3.2. Participants and results

Among the 8 participants, 6 presented dependency parsing results, and two constituency parsing results (nobody tried both subtasks). The following two Tables summarize the scores achieved by participants.

| LAS | UAS | LAS2 | Participant | Total |
|---|---|---|---|---|
| 86.94 | 90.90 | 91.59 | UniTo_Lesmo | 1-1-1 |
| 77.88 | 88.43 | 83.00 | UniPi_Attardi | 2-2-2 |
| 75.12 | 85.81 | 82.05 | IIIT_Mannem | 3-4-3 |
| 74.85 | 85.88 | 81.59 | UniStuttIMS_Schiehlen | 4-3-4 |
| * | 85.46 | * | UPenn_Champollion | *-5-* |
| 47.62 | 62.11 | 54.90 | UniRoma2_Zanzotto | 5-6-5 |

Table 3: Dependency parsing subtask evaluation.

UniTo_Lesmo achieved the best scores for dependency parsing. This rule-based parser has been developed in parallel with the TUT treebank, and so we can guess a certain influence over the annotators of the gold standard of the test set. The other parsers are statistics-based except UniRoma2_Zanzotto.

[9] http://www.di.unito.it/~tutreeb

| Br-R | Br-P | Br-F | Errors | Participant |
|---|---|---|---|---|
| 70.81 | 65.36 | 67.97 | 26 | UniNa_Corazza |
| 38.92 | 45.49 | 41.94 | 48 | FBKirst_Pianta |

Table 4: Constituency parsing subtask evaluation.

Statistics-based parsers have achieved notable results (although the development set is smaller than that in CoNLL'07), while the different tuning of the UniRoma2 Zanzotto rule-based parser can possibly explain the relatively poor performance.
For constituency format, the best result has been achieved by the UniNa_Corazza parser, again statistical parser which is an extension for Italian of Collins parser as reimplemented by Bikel.

### 3.3. Discussion

The results achieved for dependency parsing are at the state-of-the-art for Italian and very close to the state-of-the-art for English, while, as in previous experiments, those for constituency parsing are definitely far from it.
The scores of EVALITA are moreover consistent with those obtained by the application of other parsing models to TUT, and with those obtained by EVALITA participants and other parsers to the ISST[10]. The interpretation of all these results confirms that dependency parsing seems to be more adequate for the representation of Italian, as for other (relatively) free word order languages. See Bosco et al. (2008) on this same volume for a more detailed discussion.

## 4. The Word Sense Disambiguation Task

Word Sense Disambiguation (WSD) consists of associating a given word in a text or discourse with a definition or meaning.
The Senseval conferences (1998, 2001 and 2004) attempted to evaluate WSD by providing a corpus whose words had to be disambiguated according to a reference lexical resource. One of the tasks in Senseval was the all-words, in which participating systems were evaluated on their disambiguation performance on (almost) every word in the corpus.
The all-words is the task evaluated in EVALITA 2007. For each instance to disambiguate, systems have to return not only the correspondent sense(s) selected in the sense inventory of the reference resource but also its lemma and the Part of Speech (PoS) tag.

### 4.1. Data Description

The data used for the current task corresponds mostly to the set already presented in the occasion of Senseval 3 (Guazzini et al., 2004).

[10] ISST is an Italian treebank (Montemagni et al., 2003) that implements a syntactic annotation distributed on a constituent structure and a relation level including a smaller set of relations than TUT.

A corpus of about 13,600 word tokens extracted from the Italian Syntactic Semantic Treebank (Montemagni et al., 2003) was provided for testing system performance. The annotated corpus consists of a subset of 5,000 words and comprises a selection of newspaper articles about various topics. The annotation was restricted to nouns (2,583), verbs (1,858), adjectives (748), and a group of multiword expressions (97).
The reference lexical resource, provided to participants, was the ItalWordNet computational lexicon, which contains about 64,000 word senses corresponding to about 50,000 synsets (Roventini et al., 2003).

### 4.2. Evaluation Metrics

Results were evaluated by taking into account the standard measures: Precision, Recall and F-Measure (β=1). Moreover, two different scores were taken into account:

a) *Fine-grained*, in which system results are compared with the gold standard by looking for a simple correspondence.
b) *Coarse-grained*, in which an external resource (a file reporting sets of senses which can be grouped together) is used, thus allowing a more loose reckoning of the results.

### 4.3. Results and Discussion

At the beginning of the campaign, five sites registered to the task and obtained the data and guidelines. Unfortunately, at the end, only one site actually participated (Università di Bari, with the JIGSAW system). Two runs were submitted, the first containing a single guess for each token (WSD_uniba_1) and the second with multiple senses (WSD_uniba_2). Tables 5 and 6 show the results obtained by the two runs submitted by this participant regarding fine-grained and coarse-grained scores respectively.

| Run | P | R | F-measure |
|---|---|---|---|
| WSD_uniba_1 | 0.560 | 0.414 | 0.470 |
| WSD_uniba_2 | 0.503 | 0.372 | 0.427 |

Table 5: All-Words WSD results (Fine-grained)

| Run | P | R | F-measure |
|---|---|---|---|
| WSD_uniba_1 | 0.587 | 0.434 | 0.499 |
| WSD_uniba_2 | 0.519 | 0.383 | 0.440 |

Table 6: All-Words WSD results (Coarse-grained)

The participation of only one site prevents us from providing meaningful considerations about the quality of the results obtained. Nevertheless, a baseline was calculated on the basis of the "first-sense-heuristic" (in ItalWordNet the first sense is usually the commonest one) in order to introduce a term of comparison. Therefore we developed a baseline system which simply picks always the first sense. This way, we obtained quite high results (0.669 and 0.692 F-values for fine- and coarse-grained scoring respectively), in line with baselines provided within Senseval campaigns.
Finally, we would like to point out some elements of discussion that have arisen from the task:

- An element of difficulty was the fact that no training data was available for participants; the possibility of preparing training data will be considered in the event of future campaigns.
- Another point to be discussed is the complexity of a task in which systems have not only to perform WSD but also lemmatization and PoS tagging. The problem is that in this way results are less informative, since cases of incorrect PoS and lemma identification are summed to cases of incorrect disambiguation. In order to quantitatively determine the effect of PoS and lemmatization errors in the final results, we identified the errors of these types committed by the participant and re-evaluated the system without considering those tokens. The result is just a slight improvement of recall both for fine-grained (0.442 for run1 and 0.396 for run2) and for coarse-grained scoring (0.463 for run1 and 0.409 for run2). Therefore, we can state that errors due to PoS and lemmatization were not decisive on the performance.
- It is also important to mention that the participant system belongs to the non-supervised paradigm. This leads to two important considerations: on the one hand, systems of this type usually perform worse than supervised ones. On the other hand, the system could participate even if no training corpus was available, obtaining quite good results for its category if compared with results obtained in other campaigns (such as Senseval 2 and 3).

## 5. The Temporal Expression Recognition and Normalization Task

The goal of the Temporal Expression Recognition and Normalization (TERN) Task at EVALITA was to encourage research on systems capable of automatically detecting and normalizing Temporal Expressions (TEs) present in Italian texts.
Our work refers to the Automatic Content Extraction (ACE) program that in 2004 adopted the TERN Task with respect to the "TIDES 2005 Standard for the Annotation of Temporal Expressions" (Ferro et al., 2005).
TEs to be marked include both absolute (*17 luglio 2007*/July 17th, 2007) and relative expressions (*ieri*/yesterday). Also durations (*un'ora*/one hour), sets of times (*ogni settimana*/every week), underspecified expressions (*per lungo tempo*/for a long time) and TEs whose interpretation requires cultural or domain-specific knowledge (*anno accademico*/academic year) are to be annotated.
The TERN Task consisted of two subtasks based on the TIMEX2 standards with some adaptations to Italian (Magnini et al., 2007a): (i) Temporal Expression Recognition only, in which systems are required to recognize the TEs occurring in the source data by identifying their extension; (ii) Temporal Expression

Recognition + Normalization, in which systems are required to give a representation of the meaning of TEs by assigning values to a pre-defined set of attributes.

### 5.1 Data Description

Both training data and test data are part of the Italian Content Annotation Bank (I-CAB), developed by FBK and CELCT (Magnini et al., 2006).
I-CAB consists of 525 news stories taken from different sections (e.g. Cultural, Economic, Sports and Local News) of the local newspaper "L'Adige", for a total of around 180,000 words (the ratio between training and test data was 2/1). The total number of annotated TEs is 4,603: 2,931 and 1,672 in the training and test sections respectively.
The manual annotation of the corpus was rather time-consuming: the realization of a gold standard with the possible minimum number of inconsistencies and errors, in fact, required 1 person/year.
I-CAB version 4.1, used in EVALITA, is freely available for research purposes[11].

### 5.2 Evaluation metrics and results

The final ranking is based on the TERN value score, already adopted in the ACE program. The value score is defined to be the sum of the values of all of the system's output TIMEX2 tokens, normalized by the sum of the values of all of the reference TIMEX2 tokens.
We also provided the Precision, Recall and F-Measure.
Tables 7 and 8 present the results for both tasks in term of TERN-Value score, Precision (P), Recall (R) and F-measure (F).

| Participant | Value | P | R | F |
|---|---|---|---|---|
| FBKirst_Negri_TIME | **85.7** | 95.7 | 89.8 | 92.6 |
| UniPg_Faina_TIME | **50.1** | 77.7 | 70.3 | 73.8 |
| UniAli_Puchol_TIME | **48.8** | 78.4 | 67.4 | 72.5 |
| UniAli_Saquete_TIME | **41.9** | 82.5 | 53.2 | 64.7 |

Table 7: Results for the Recognition only subtask, percentages for Value, Precision, Recall and F-measure

| Participant | Value | P | R | F |
|---|---|---|---|---|
| FBKirst_Negri_TIME | **61.9** | 68.5 | 63.3 | 67.4 |
| UniAli_Saquete_TIME | **22.1** | 51.5 | 35.6 | 42.1 |
| UniPg_Faina_TIME | **11.9** | 24.9 | 19.6 | 21.9 |

Table 8: Results for the Recognition + Normalization subtask, percentages for Value, Precision, Recall and F-measure

The Value scores achieved by participant systems ranged from 41.9% to 85.7% in the Recognition only subtask, while, for the Recognition + Normalization subtask, the systems obtained between 11.9% and 61.9%. The submissions of FBKirst_Negri_TIME stand out as more than 35% higher than the other systems in both the task.

[11] http://tcc.itc.it/projects/ontotext/i-cab/download-icab.htm

### 5.3 Discussion

Four teams participated in the challenge: three in the Recognition + Normalization subtask and one in the Recognition only subtask. FBKirst_Negri and UniAli_Saquete systems adopt a rule-based approach in both the subtasks, while UniAli_Puchol participated to the Recognition only subtask with a machine learning system. Finally, the UniPg_Faina system is a parser with a good result in the Recognition only subtask but with a very low value score in the Normalization subtask.
We appreciated the participation of two foreign groups to the task: they both extended to Italian their original systems developed for Spanish, using an automatic translation of the existing temporal models.
We received the expected attention in terms of participation: actually, eight groups registered but four of them could not adjust their system in time. Considering that this was a new and relatively difficult task for the Italian language, this is quite understandable. We hope that the number of participants will grow in the next evaluation campaigns. The TERN Task, indeed, is a key step in the Information Extraction field so it's necessary that the research community, in particular the Italian one, invests more in this field.

## 6. The Named Entity Recognition Task

The Named Entity Recognition (NER) Task evaluated system performance at recognizing four different types of Named Entities, i.e. Person (PER), Organization (ORG), Geo-Political Entity (GPE) and Location (LOC). The task was based on the ACE-LDC standards for the ACE Entity Recognition and Normalization Task[12], with appropriate adaptations needed to limit it to the recognition of Named Entities (NEs) only (Magnini et al. 2007b).

### 6.1. Data Description and Evaluation Metrics

As a dataset, we used the I-CAB corpus, developed within the Ontotext project and described in Section 5.1. Training and test data contained respectively 7,434 and 3,976 NEs. PER was the most frequent type of NEs (40% of the total), followed by ORG (32%), GPE (25%), and LOC (only 3%).
Participants were provided with training data annotated in the IOB2 format, where every token was annotated with a tag: 'B' ('begin') for the first token of each NE, 'I' ('inside') for other tokens of the NE, and 'O' ('outside') for tokens that did not belong to any NE; tags 'B' and 'I' were followed by the NE type.
Inter-annotator agreement was evaluated on the dual annotation of a subset of the corpus using the Dice coefficient (computed as Dice=2C/(A+B), where C is the number of common annotations, while A and B are the number of annotations provided by the two annotators). The values of the Dice coefficient we obtained were quite high: 96% for PER, 84% for ORG, 97% for GPE and 89% for LOC Entities.
The NER Task at EVALITA 2007 had six participants

[12] ACE Program: http://projects.ldc.upenn.edu/ace

from four different countries: University of Alicante and Yahoo! from Spain, University of Dortmund and University of Duisburg-Essen from Germany, LDC from the USA, and Fondazione Bruno Kessler from Italy.
For the official evaluation of system results we used the scorer made available for the CoNLL-2002 Shared Task[13]. System results (each participant was allowed to submit up to two runs) were evaluated using standard measures, i.e. Precision (the ratio between the number of NEs correctly identified and the total number of NEs identified) and Recall (the ratio between the number of NEs correctly identified and the number of NEs that the system was expected to recognize); the official ranking was based on the F-Measure, i.e. the weighted harmonic mean of Precision and Recall.

### 6.2 Results and Discussion

The F-Measure values achieved by participants (see Table 9 for the best run of each participant) ranged between 82.14 and 63.10, with half of them between 66 and 69, and two above 70. System results have been compared with two different baseline rates computed by identifying in the test data only the NEs that appeared in the training data. In one case, NEs which had more than one type in the training data were not taken into consideration (FB1=36.85); in the other case, they were annotated with the most frequent type (FB1=41.11). As far as Precision and Recall are concerned, most systems obtained higher values for Precision than for Recall, with only two exceptions.

| Participant | FB1 | Prec. | Recall |
|---|---|---|---|
| FBKirst_Zanoli_r2 | **82.14** | 83.41% | 80.91% |
| UniDuE_Roessler_r1 | **72.27** | 71.62% | 72.94% |
| Yahoo_Ciaramita_r1 | **68.99** | 71.28% | 66.85% |
| UniDort_Jungermann_r2 | **67.90** | 70.93% | 65.12% |
| UniAli_Kozareva | **66.59** | 62.73% | 70.95% |
| LDC_Walker_r1 | **63.10** | 83.05% | 50.88% |
| BASELINE | **41.11** | 42.44% | 39.86% |
| BASELINE -u | **36.85** | 40.29% | 33.95% |

Table 9: Results of the NER Task at EVALITA 2007.

In spite of the differences between the CoNLL-2003 Shared Task on language-independent NER (Tjong Kim Sang & De Meulder 2003) and the NER Task at EVALITA 2007 (in the first place, the different types of NEs to be recognised), it is still interesting to compare the results of the two evaluations. The best system at EVALITA 2007, in fact, scored slightly lower than the best system for English in the Shared Task (which scored 88.76); the results of the second best system, on the other hand, are very close to the performance of the best system for German in the Shared Task (which scored 72.41).
The highest scores at EVALITA 2007 were obtained by FBKirst_Zanoli and UniDuE_Roessler with two machine learning systems exploiting Support Vector Machines (EntityPro and Walu, respectively). The most significant difference between the two systems is that EntityPro, unlike WALU, was enriched with gazetteers and other external resources, which partly explains the ten-point difference in their results. As reported by Pianta and Zanoli (2007), in fact, the performance of EntityPro drops by about eight points when used without external resources.
The recognition of PER NEs turned out to be the easiest subtask, as all participants obtained their highest F-Measure values, ranging from 75 to 92 (Speranza, 2007). The recognition of NEs of type GPE did not constitute a problem for most participant systems either, with F-Measure values ranging between 65 and 86. System results dropped significantly in the recognition of NEs of type LOC , ranging between 46 and 73; the effect of such results on the overall performance of the systems, however, was limited by the low frequency of LOC NEs in the corpus. The most problematic subtask was undoubtedly the recognition of NEs of type ORG, where all systems except one obtained their lowest results, none being able to perform better than 65.
With the participation of six institutions from four different countries, we feel that we have achieved our initial goal of fostering research on Named Entity Recognition for Italian although we had only one Italian institution among our participants. We hope that the outcome of the NER Task at EVALITA 2007 will help stimulate the organization of further evaluation campaigns in the field of NER for Italian, where it might be interesting to propose more complex tasks, such as the identification of entity attributes and co-reference, in addition to the basic NER Task.

## 7. Conclusions

The application of existing methods to different languages and data sets is crucial, since the validation of existing NLP models strongly depends on the possibility of generalizing their results on data and languages other than those on which they have been trained and tested. Therefore, establishing shared standards, resources, tasks and evaluation practices with reference to languages other than English is a fundamental step towards the continued development of NLP.
The EVALITA experience can be seen as the first picture of the problems that lie ahead for Italian NLP and the kind of work necessary for adapting existing models to this language, both in terms of systems and resources.
In fact, on the one hand, the good response obtained by this initiative, both in the number of participants and in the quality of results, often near the state-of-the-art, showed that it is worth pursuing such goals for Italian. On the other hand, this event has given us a clearer assessment of both the distribution of NLP research groups in Italy and for Italian, and the complexity of proposed tasks also with reference to the state of development of Italian linguistic resources.
As an immediate effect, thanks to the cooperation

[13] Freely available at: http://www.cnts.ua.ac.be/conll2002/ner/

between organizers and participants, the evaluation campaign resulted in an increased amount of training and test data compliant with international standards, as well as being more reliable than previously, which have been made available to the scientific community and remain as benchmarks for future improvements.

## 8. Acknowledgements

This work was partially supported by the three-year project Ontotext, funded by the Autonomous Province of Trento.

## 9. References

Adda G., Lecomte J., Mariani J., Paroubek P., Rajman M. (1998). The GRACE French Part-of-Speech Tagging Evaluation Task. In *Proc. LREC'98*.

Bosco, C., Mazzei, A., Lombardo, V., Attardi, G., Corazza, A., Lavelli, A., Lesmo, L., Satta, G., Simi, M. (2008) Comparing Italian parsers on a common treebank: the EVALITA experience. In *Proc. LREC'08*.

Bosco, C., Mazzei, A., Lombardo, V. (2007) EVALITA parsing task: an analysis of the first parsing system contest for Italian. *Intelligenza Artificiale*, 4(2).

Bosco C., Lombardo V., Vassallo D., Lesmo L. (2000). Building a treebank for Italian: a data-driven annotation schema. In *Proc. LREC'2000*.

Brants T. (2000). TnT - A Statistical Part-of-Speech Tagger. In *Proc. 6th Conf. on Applied Natural Language Processing*, pp. 224-231.

Brill E. (1994). Some Advances in Transformation-Based Part of Speech Tagging. In *Proc. 12th National Conference on Artificial Intelligence*, pp. 722-727.

Collins, M., Hajic, J., Ramshaw, L. and Tillmann, C. (1999) A statistical parser of Czech. In *Proc. of ACL'99*.

Corazza, A., Lavelli, A., Satta, G. and Zanoli, R. (2004). Analyzing an Italian treebank with state-of-the-art statistical parser. In *Proc. of TLT-2004*.

Daelemans W., Zavrel J., Berck S. (1996). MBT: A Memory Based Part of Speech Tagger-Generator. In *Proc. 4th Workshop on Very Large Corpora*, pp. 14- 27.

Delmonte R. (2004). Strutture sintattiche dall'analisi computazionale di corpora di italiano. In A. Cardinaletti, A. and F. Frasnedi (Eds.), *Intorno all'italiano contemporaneo. Tra linguistica e didattica*. F. Angeli, Milano, pp. 187-220.

Ferro, L., Gerber, L., Mani, I., Sundheim, B., and Wilson, G. (2005). TIDES 2005 Standard for the Annotation of Temporal Expressions. http://timex2.mitre.org/annotation_guidelines/2005_timex2_standard_v1.1.pdf.

Gildea, G. (2001). Corpus variation and parser performance. In *Proc. of EMNLP'01*.

Guazzini, E., Ulivieri, M., Bertagna, F., Calzolari., N., (2004). Senseval-3: the Italian All-Words Task. In *Proc. SENSEVAL-3*.

Magnini, B., Cappelli, A. (Eds.) (2007). Proc. of EVALITA 2007. *Intelligenza Artificiale*, 4(2).

Magnini, B., Negri, M., Pianta, E., Speranza, M., Bartalesi Lenzi, V., and Sprugnoli, R. (2007a). Italian Content Annotation Bank (I-CAB): Temporal Expressions. Technical Report, FBK-irst. http://evalita.itc.it/tasks/I-CAB-Report-Temporal-Expressions.pdf.

Magnini, B., Pianta, E., Speranza, M., Bartalesi Lenzi, V., and Sprugnoli, R. (2007b). Italian Content Annotation Bank (I-CAB): Named Entities, Technical Report FBK-irst. http://evalita.itc.it/tasks/I-CAB-Report-Named-Entities.pdf.

Magnini, B., Cappelli, A., Pianta, E., Speranza, M., Bartalesi Lenzi, V., Sprugnoli, R., Romano, L., Girardi C., Negri, M. (2006). Annotazione di contenuti concettuali in un corpus italiano: I-CAB. In *Proc. of SILFI 2006*, Florence, Italy.

Monachini M. (1996). *ELM-IT: EAGLES Specification for Italian morphosintax Lexicon Specification and Classification Guidelines*. EAGLES Document EAG CLWG ELM IT/F.

Montemagni, S., Francesco, F., Battista, M., Calzolari, N., Corazzari, O, Lenci, A., Pirrelli, V., Zampolli, A., Fanciulli, F., Massetani, M., Raffaelli, R., Basili, R., Pazienza, M.T., Saracino, D., Zanzotto, F., Mana, N., Pianesi, F., Delmonte, R. (2003). The syntactic-semantic Treebank of Italian. An Overview. *Linguistica Computazionale* a Pisa, vol. I.

Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S. and Yuret, D. (2007). The CoNLL-2007 shared task on dependency parsing. In *Proc. of EMNLP- CoNLL*.

Pianta, E., Zanoli, R. (2007). Exploiting SVM for Italian Named Entity Recognition. In Proc. of EVALITA 2007. *Intelligenza Artificiale*, 4(2).

Ratnaparkhi A. (1996). A Maximum Entropy Model for Part-of-Speech Tagging. In *Proc. EMNLP'96*.

Roventini, A. Alonge, A., Bertagna, F., Calzolari, N., Girardi, C., Magnini, B., Marinelli, R., Speranza, M., Zampolli, A. (2003). ItalWordNet: Building a Large Semantic Database for the Automatic Treatment of Italian. In *Linguistica Computazionale*, Istituti Editoriali e Poligrafici Internazionali, Pisa-Roma, ISSN 0392-6907.

Speranza, M. (2007). EVALITA 2007: The Named Entity Recognition task. In Proc. of EVALITA 2007. *Intelligenza Artificiale*, 4(2).

Tamburini F., Seidenari C. (2007). *EVALITA 2007. The Italian Part-of-Speech Tagging Evaluation - Task Guidelines*. http://evalita.itc.it/tasks/pos.html.

Tjong Kim Sang, E., De Meulder, F. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. *Proc. of CoNLL-2003 at HLT-NAACL*, Edmonton, Canada, 2003.